Selecting Sketches for Similarity Search

Vladimir Mic[1], David Novak[1], Lucia Vadicamo[2] and Pavel Zezula[1]

[1]Masaryk University, Brno, Czech Republic, [2]CNR-ISTI, Pisa, Italy

Abstract. Techniques of the Hamming embedding, producing bit string sketches, have been recently successfully applied to speed up similarity search. Sketches are usually compared by the Hamming distance, and applied to filter out non-relevant objects during the query evaluation. As several sketching techniques exist and each can produce sketches with different lengths, it is hard to select a proper configuration for a particular dataset. We assume that the (dis)similarity of objects is expressed by an arbitrary metric function, and we propose a way to efficiently estimate the quality of sketches using just a small sample set of data. Our approach is based on a probabilistic analysis of sketches which describes how separated are objects after projection to the Hamming space.

1 Introduction

Efficient search for data objects according to their pairwise similarity presents an important task in data processing. We consider similarity search for complex objects, e.g. multimedia files. Features of these objects are typically characterized by *descriptors*, which are often high dimensional vectors. They can be bulky and evaluation of their pairwise similarity may be computationally demanding. Thus techniques to process them efficiently are needed. We consider one to one mapping between objects and descriptors, thus we do not distinguish these terms and we use just term object.

Techniques transforming data objects to smaller objects are often used to speed up similarity search. Their number and number of their inherent parameters make their fair comparison difficult. Moreover, the ability of particular approaches to approximate similarity relationships between objects is data dependent. This paper considers a particular family of techniques – transformation of objects to the Hamming space – and it provides formal analysis which allows to efficiently estimate a quality of particular transformation techniques. Our approach uses a small sample set of the original and the transformed objects, and, inspired by the *separability*, which is traditionally used in data clustering and processing of biometrics, we estimate the ability of the transformed objects to distinguish different values of genuine similarity. We define the problem precisely in the following section.

1.1 Problem Formulation

We focus on similarity search in the *metric space* [18]. The notation used throughout the paper is provided in Table **??**. Formally, the metric space is a pair (D, d)

Table 1: Notation used throughout this paper

(D, d); $X \subseteq D$	metric space with domain D and distance function d; dataset X
$iDim$	intrinsic dimensionality of dataset
$sk(o)$	sketch of object $o \in X$
λ	length of sketches in bits
$h(sk(o_1), sk(o_2))$	Hamming distance of sketches $sk(o_1)$ and $sk(o_2)$
β	balance of bits of sketches
$p(x, b)$	probability that $h(sk(o_1), sk(o_2)) = b$ for $o_1, o_2 \in X : d(o_1, o_2) = x$
$p_i(x, 1)$	probability $p(x, b)$ for $\lambda = 1$ considering an average bit i
x, b	values of the distance functions d and h, respectively
Γ	maximum distance x
ϕ	number of degrees of freedom of distance function $d(o_1, o_2)$
μ, σ^2	mean value and variance of Hamming distance
m, s^2	mean value and variance of probability $p(x, b)$ for a given x

where D is a domain of objects and d is a total *distance function* $d : D \times D \mapsto \mathbb{R}$. This function determines the dissimilarity of objects – the bigger the value $d(o_1, o_2)$, the less similar the objects $o_1, o_2 \in D$. The distance can be an arbitrary function which satisfies the properties of non-negativity, identity, symmetry and triangle inequality [16].

Having a metric space (D, d), we consider a dataset $X \subseteq D$, and a *sketching technique* $sk : D \mapsto \{0, 1\}^\lambda$, which transforms objects $o \in D$ to bit-strings of fixed length λ. We call these bit strings *sketches*, and we assume that dissimilarity of these sketches is measured by the *Hamming distance*. Further, we focus on a family of sketching techniques which produce bits *balanced to* β:

- Bit i is balanced to ratio β (with respect to the dataset X) iff it is set to 1 in $\beta \cdot |X|$ sketches $sk(o), o \in X$.

We consider just values $0.5 \leq \beta \leq 1$, since if β is smaller than 0.5 for some bit i, this bit can be flipped in all sketches $sk(o), o \in X$ which preserves all the Hamming distances. The objective of this paper is to propose a way to estimate ability of sketches to approximate similarity relationships between objects $o \in X$, using just a small sample set of data and sketches.

1.2 Related Work

Several sketching techniques have been proposed [1, 5, 7, 10–14, 17] and most of them produce sketches with bits balanced to 0.5. To the best of our knowledge, there is no prior work which would efficiently estimate, what sketches, their balance β and length λ are suitable for particular data (on condition that β is tunable). For instance, Wang et al. [17] provide analysis to estimate recall of kNN queries for particular sketching technique to select suitable length of sketches. However, their method is not extendible for other sketching techniques and

the estimation is rather approximate. Mic et al. also provide approach to estimate suitable length λ for their sketching technique [12], but their ideas cannot be applied for arbitrary sketches. Our proposal is inspired by Daugman's analysis [3], who investigates binary codes of human irises. He evaluates *separability* of two distance densities to describe the quality of his method to identify people according to this biometric.

The paper is organized as follows. Section 2 contains analysis to estimate an ability of sketches to approximate similarity relationships between objects, Section 3 proposes another approach to estimate this quality, Section 4 contains discussion about a cost of estimations and results of experiments to compare measured and estimated quality of sketches, and Section 5 concludes the paper.

2 Analysis to Estimate Quality of Sketches

The goal of this section is to derive formula describing the ability of sketches to separate two distances from the original metric space. In particular, we consider four arbitrarily selected objects $o_1, o_2, o_3, o_4 \in X$ and their distances $d(o_1, o_2) = x_1, d(o_3, o_4) = x_2, x_1 \leq x_2$. The goal of function $sep_{sk}(x_1, x_2)$ is to describe how separated are the Hamming distances $h(sk(o_1), sk(o_2))$ and $h(sk(o_3), sk(o_4))$.

2.1 A Single-bit Sketch

We start with an average probability $p_i(x, 1)$ that one bit i of sketches $sk(o_1)$ and $sk(o_2)$ has different value for objects $o_1, o_2 \in X$ with distance $d(o_1, o_2) = x$. This probability can be derived in an analytic way just for some specific sketching techniques [17]. Therefore, we propose to determine it empirically by its evaluation on a sample set of data. We measure the probability $p_i(x, 1)$ on equidistant intervals of distances x. To make function $p_i(x, 1)$ continuous, we use linear interpolation between measured points and we add an artificial point [0, 0] to catch influence of smaller distances than were observed on the sample set. We work with an average probability $p_i(x, 1)$ evaluated over all bits i. Probability function $p_i(x, 1)$ constitutes one of features describing quality of sketches, as it should obviously increase with x. An example is provided in Figure 1a.

2.2 Projection of Distance x on Hamming Distance b

As a next step, we derive probability function $p(x, b)$ that Hamming distance $h(sk(o_1), sk(o_2))$ is equal to b for objects o_1, o_2 with distance $d(o_1, o_2) = x$. It is done by composition of λ instances of probability function $p_i(x, 1)$. This step is challenging due to possible bit correlations. Probability function $p(x, b)$ for a fixed x can be modelled by a symmetric function[1], which allows us to use its *binomial analogue*. It is a scaled binomial distribution with the same variance as function $p(x, b)$. To fit variance of function $p(x, b)$, we need to estimate its *number of degrees of freedom* ϕ [3].

[1] Reasoning is provided at https://www.fi.muni.cz/ xmic/sketches/Symmetry.pdf

Lemma 1. *The number of degrees of freedom ϕ of function $p(x, b)$ is similar to the number of degrees of freedom ϕ' of the density of the Hamming distance on all sketches $sk(o), o \in X$.*

Clarification Daugman [3] evaluates the number of degrees of freedom of the density of the Hamming distance on sketches:

$$\phi' = \frac{\mu \cdot (\lambda - \mu)}{\sigma^2} \tag{1}$$

where λ is the length of sketches, μ is the mean value and σ^2 is the variance of the Hamming distance. According to analysis in [13], the μ is given by λ and β, and σ^2 is given by λ, β and pairwise bit correlations. Therefore, iff sketches of objects $o_1, o_2 : d(o_1, o_2) = x$ have bits balanced to β and they have same pairwise bit correlations as all the sketches $sk(o), o \in X$, the Lemma 1 describes equality, i.e. $\phi = \phi'$. Our first approach to estimate quality of sketches assumes this equation, and the error caused by this assumption is discussed in Section 2.4.

We connect Equation 1 with the term *intrinsic dimensionality* (*iDim*), which describes an amount of information in data. Several ways to estimate the *iDim* have been developed but just a few of them can be used in a general metric space. We use the formula of Chávez and Navarro [2]:

$$iDim \approx \frac{\mu^2}{2 \cdot \sigma^2}. \tag{2}$$

The mean value μ equals $2\lambda \cdot \beta \cdot (1 - \beta)$ [13], and thus, using the Equations 1, 2 and Lemma 1, we may express the number of degrees of freedom ϕ using the intrinsic dimensionality of sketches *iDim* and balance of their bits β:

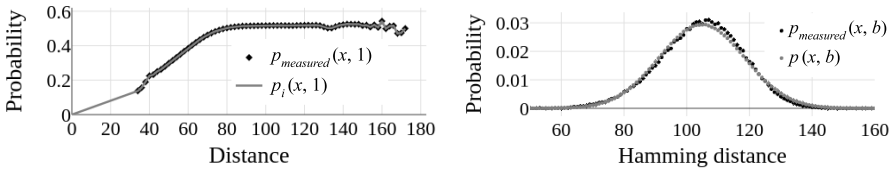
$$\phi = \frac{\mu \cdot (\lambda - \mu)}{\sigma^2} = \frac{\mu \cdot \frac{\mu}{2 \cdot \beta \cdot (1-\beta)}}{\sigma^2} - \frac{\mu^2}{\sigma^2} \approx 2 \cdot iDim \cdot \left(\frac{1}{2 \cdot \beta \cdot (1 - \beta)} - 1 \right). \tag{3}$$

In order to model probability $p(x, b)$, we propose to use binomial distribution with ϕ degrees of freedom which we scale and interpolate to get the final function. The only input necessary for the usage of this binomial analogue is *iDim* of sketches, empirically evaluated on a sample set of sketches, and balance of their bits β. We round number of degrees of freedom to the nearest integer and denote it ϕ in the rest of this paper. In the following, we describe the estimation of $p(x, b)$ formally. We approximate this function by a linear interpolation $p_{\text{lin}}(x, b)$ normalized with a coefficient *coef*(x):

$$p(x, b) \approx \frac{p_{\text{lin}}(x, b)}{coef(x)}. \tag{4}$$

The normalization coefficient *coef*(x) is evaluated as:

$$coef(x) = \sum_{i=0}^{\phi} p_{\text{lin}}(x, i)$$



(a) Dependency of $p_i(x, 1)$ on x (b) Measured and modelled $p(x, b)$

Fig. 1: Example of functions $p_i(x, 1)$ and $p(x, b)$ used in the model

and the linear interpolation $p_{\text{lin}}(x, b)$ as:

$$p_{\text{lin}}(x, b) = p_{\text{int}}(x, \lfloor b' \rfloor) + (b' - \lfloor b' \rfloor) \cdot (p_{\text{int}}(x, \lceil b' \rceil) - p_{\text{int}}(x, \lfloor b' \rfloor))$$

where b' is the scaled value b:

$$b' = b \cdot \frac{\phi}{\lambda}$$

and where function $p_{\text{int}}(x, b_{\text{int}})$ (which requires the second parameter to be an integer), evaluates the binomial distribution:

$$p_{\text{int}}(x, b_{\text{int}}) = \binom{\phi}{b_{\text{int}}} \cdot p_i(x, 1)^{b_{\text{int}}} \cdot (1 - p_i(x, 1))^{\phi - b_{\text{int}}}.$$

We use linear interpolation p_{lin} since b' is usually not an integer. Due to the transformation of the binomial distribution, it is necessary to normalize probability using *coef*(x) coefficient: we normalize function p_{int} by the sum over all values, since $p(x, b)$ is discrete with respect to b.

We show an example, how this binomial analogue fits the distribution of the measured $p(x, b)$ in Figure 1b. The black points show values empirically measured, and grey points show values estimated by the proposed binomial analogue described by Equation 4. In this experiment, we used sketches with $\lambda = 205$ bits balanced to $\beta = 0.5$, and just sketches of objects within distance $x = 86$ evoking probability $p_i(x, 1) = 0.51$. These measurements confirm that the binomial analogue is a good approximation of probability $p(x, b)$.

2.3 Quality of Sketches

Quality of sketches used for the similarity search is given by their ability to preserve similarity relationships between objects. Let us consider four arbitrarily selected objects $o_1, o_2, o_3, o_4 \in X$ and distances $d(o_1, o_2) = x_1$ and $d(o_3, o_4) = x_2, x_1 \leq x_2$. In the following, we focus on a *separation* of probability functions $p(x_1, b)$ and $p(x_2, b)$ and we describe it by formula adopted from work of Daugman [3]:

$$sep_{sk}(x_1, x_2) = \frac{m_2 - m_1}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \tag{5}$$

where m_1 and m_2 are mean values of $p(x_1, b)$ and $p(x_2, b)$, and s_1^2 and s_2^2 are their variances. These values can be expressed by analysis of Equation 4:

Lemma 2. *Mean value m of probability function $p(x, b)$ is $p_i(x, 1) \cdot \lambda$. Variance s^2 of probability function $p(x, b)$ is $\frac{\lambda^2}{\phi} \cdot p_i(x, 1) \cdot (1 - p_i(x, 1))$.*

Proof. Function $p(x, b)$ is formed by binomial distribution (see function p_{int} in Equation 4), which is scaled with respect to value b by coefficient $\frac{\phi}{\lambda}$. Since mean of function p_{int} is $p_i(x, 1) \cdot \phi$ and its variance is $\phi \cdot p_i(x, 1) \cdot (1 - p_i(x, 1))$, then:

$$m = p_i(x, 1) \cdot \phi \cdot \frac{\lambda}{\phi} = p_i(x, 1) \cdot \lambda,$$

and

$$s^2 = \phi \cdot p_i(x, 1) \cdot (1 - p_i(x, 1)) \cdot \left(\frac{\lambda}{\phi}\right)^2 = \frac{\lambda^2}{\phi} \cdot p_i(x, 1) \cdot (1 - p_i(x, 1)).$$

Theorem 1. *Considering four arbitrary objects $o_z \in X, z \in [1..4]$ with distances $d(o_1, o_2) = x_1$, $d(o_3, o_4) = x_2, x_1 \leq x_2$, and an arbitrary sketching technique sk producing sketches with bits balanced to β, the separation $sep_{sk}(x_1, x_2)$ of the Hamming distances $h(sk(o_1), sk(o_2))$ and $h(sk(o_3), sk(o_4))$ can be expressed:*

$$sep_{sk}(x_1, x_2) \approx 2 \cdot \sqrt{iDim} \cdot f_{sk}(x_1, x_2) \cdot \sqrt{\frac{1}{2 \cdot \beta \cdot (1 - \beta)} - 1} \tag{6}$$

where

$$f_{sk}(x_1, x_2) = \frac{p_i(x_2, 1) - p_i(x_1, 1)}{\sqrt{p_i(x_1, 1) \cdot (1 - p_i(x_1, 1)) + p_i(x_2, 1) \cdot (1 - p_i(x_2, 1))}} \tag{7}$$

Proof. Theorem holds as a consequence of Equation 3, Equation 5, and Lemma 2.

Theorem 1 reveals features of sketches $sk(o), o \in X$, which improves their capability to approximate similarity relationships between objects. For instance, sufficiently high *iDim* of sketches is necessary to allow them distinguish distances $d(o_1, o_2) < d(o_3, o_4)$. Please notice, that just function $f_{sk}(x_1, x_2)$ (defined by Equation 7) takes into account values x_1 and x_2, and that there is no direct dependency of $sep_{sk}(x_1, x_2)$ on sketch length λ.

To describe quality of sketches, we propose to evaluate $sep_{sk}(x_1, x_2)$ over whole range of distances function d. Without loss of generality, we assume that d is continuous and its range is $[0, \Gamma]$. Then:

$$quality(sk) = \int_0^{\Gamma} \int_{x_1}^{\Gamma} sep_{sk}(x_1, x_2)\, \partial x_2\, \partial x_1 \tag{8}$$

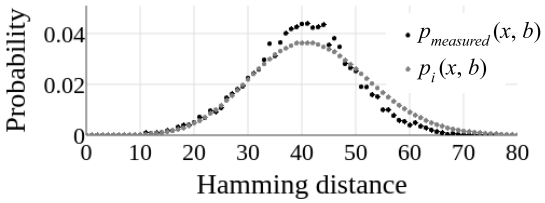


Fig. 2: Measured and modelled $p(x, b)$ for $x = 38$ implying $p_i(x, 1) = 0.2$

Semantics of this integral is in compliance with the sign of $sep_{sk}(x_1, x_2)$: value of $sep_{sk}(x_1, x_2)$ is negative iff the distances x_1, x_2 are swapped after the transformation to sketches[2]. Such distances x_1, x_2 then naturally decrease the quality of sketching technique, as described by this equation. Since $quality(sk)$ cannot be meaningfully compared for metrics with different Γ, we propose to normalize it. Equation 8 allows direct normalization by Γ^2. Finally, if sketches are going to be applied for similarity search, they are needed to well separate small distances from the others. Therefore, it is meaningful to evaluate quality of sketching technique using some threshold $t < \Gamma$:

$$quality_{norm}(sk, t) = \frac{\int_0^t \int_{x_1}^{\Gamma} sep_{sk}(x_1, x_2)\, \partial x_2\, \partial x_1}{\Gamma^2}. \tag{9}$$

Therefore, we propose to evaluate $quality_{norm}(sk, t)$ using a sample set of data and use it to compare quality of sketches. The cost of this estimation is discussed in Section 4.1.

2.4 Sources of Error

The main source of error of proposed quality estimation is caused by Lemma 1, as we assume that balance β and pairwise correlations of bits of sketches $sk(o_1)$, $sk(o_2), o_1, o_2 \in X : d(o_1, o_2) = x$ for an arbitrary given x are the same, as on the whole dataset X. The precision of this assumption is data and sketching technique dependent. We have observed, that the proposed binomial analogue is quite precise for non-extreme distances x (as shown by Figure 1b), but its precision decreases mainly for tails of x. Therefore, this feature causes an erroneous estimation for some sketching techniques and datasets. An example is given in Figure 2, where the binomial analogue $p(x, b)$ for a very low x is examined.

Another errors are caused by our intention to use a small sample set to evaluate $quality_{norm}(sk, t)$, and by the fact that the provided analysis is based on expected values and it does not consider deviations from modelled functions. We evaluate experiments with this approach in Section 4, and we denote it *Approach A* (as *analytique)*. In the following section, we propose the second way to estimate quality of sketches, which aims to mitigate the error of Approach A.

[2] Please see, that the sign of $sep_{sk}(x_1, x_2)$ is given by the sign of function $f_{sk}(x_1, x_2)$, and this is negative iff $p_i(x_2, 1) < p_i(x_1, 1)$. We have assumed $x_1 \leq x_2$, and these two inequalities are equivalent to swapping distances x_1, x_2.

3 Approach PM

The second way to estimate $quality_{norm}(sk, t)$ is based on direct evaluation of means m_1, m_2 and variances s_1^2, s_2^2 of the $p(x, b)$, used in Equation 5. If these values are evaluated directly on a sample set of data, just this equation and Equation 9 are utilized to estimate $quality_{norm}(sk, t)$. However, this approach requires evaluation of mean m and variance s^2 of $p(x, b)$ for the whole range of distances $x \in [0..\Gamma]$.

In particular, we use equidistant intervals of x and we evaluate distances $b = h(sk(o_1), sk(o_2))$ for each pair of objects o_1, o_2 from the sample set such that $d(o_1, o_2)$ is from a given interval of x. Then we evaluate the mean m and variance s^2 for each interval of x and we add an artificial mean $m = 0$ and variance $s^2 = 0$ for distance $x = 0$. Finally, we use linear interpolation to get values m and s^2 for an arbitrary distance x.

We denote the estimation of $quality_{norm}(sk, t)$ according to this procedure *Approach PM* (*partially measured*) and we evaluate its capability to estimate quality of sketches in Section 4. At first, let us discuss sources of errors of this approach. The cost of this estimation is discussed in Section 4.1.

3.1 Sources of Error of the PM Approach

Approach PM mitigates an error brought to the Approach A by too strong usage of Lemma 1. Instead of this error, Approach PM is more sensitive on a low number of objects o_1, o_2 within very small distances $d(o_1, o_2) = x$ in the sample set. Since the mean m and variance s^2 of $p(x, b)$ is examined for the whole range of distance x, it needs a representative number of objects o_1, o_2 within each interval of x to measure it precisely, which is obviously a problem for tails of distances x. However, exactly the ability to precisely handle extremely small distances is crucial to well estimate quality of sketches for similarity search.

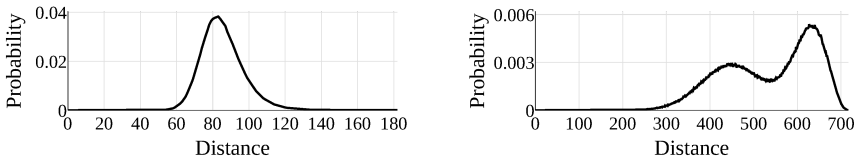
The rest of errors is caused by similar features as in case of Approach A (see Section 2.4 for details). Therefore, both approaches provide estimation with some probable level of error, and we evaluate them both in the next section.

4 Experiments

This section provides verification of the proposed approaches to estimate quality of sketches. At first, we discuss costs of proposed estimations in comparison with a traditional approach to evaluate quality of sketches.

4.1 Queries Evaluations vs. Proposed Estimations

Testing quality of sketches is usually performed via evaluation of sufficient number of representative queries on sample data. Therefore, the precise answer for these queries must be known, and the sketches for both, the dataset and query set must be created.



(a) DeCAF (b) SIFT

Fig. 3: Distance densities of $d(o_1, o_2), o_1, o_2 \in X$

We use the recall of k nearest neighbours queries (kNN) evaluated via simple sketch-based filtering. In particular, sketches are applied to *filter* a fixed number of the most similar sketches to the query sketch $sk(q)$, and then the set of corresponding objects $CandSet(q)$ is *refined* by the distances $d(q, o), o \in CandSet(q)$. Finally, the k most similar objects from the $CandSet(q)$ are returned as the query answer. The *recall* expresses the relative intersection of this answer with the precise answer returned by the sequential evaluation of all distances $d(q, o), o \in X$. Please, notice that a suitable size of the $CandSet(q)$ considering applied sketching technique must be selected, which is another difficult and data dependent task. Finally, even if this expensive procedure is performed, it is relevant just for a tested dataset and it is dependent on a selection of queries.

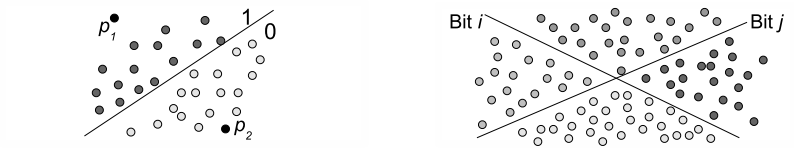
We evaluate one thousand 100NN queries on two datasets of size 1 million objects. Therefore, both ground truths cost 2 billion evaluations of distance $d(q, o)$, several (30) different sets of million sketches are created, and finally, billion Hamming distances are evaluated for each set of sketches. In our experiments, we use 16 and 14 different sets of sketches for our datasets. The cost of their creation is in order of billions of distance computations, and several GB of data are read from hard-drives during these experiments.

Conversely, proposed quality estimations do not use any queries. We use just a sample set of 5000 objects and their sketches in our experiments. In case of Approach A, we use 2 million distances to get function $p_i(x, 1)$ and $iDim$ of sketches. The efficiency of the estimation is given mainly by the precision of integral evaluation (defined by Equation 9). Since we use parameters providing high precision, an evaluation of $quality_{norm}(sk, t)$ by Approach A for one set of sketches takes about 50 seconds on average. Approach PM is even more efficient, as it uses 2 millions distances to get means m and variances s^2 of $p(x, b)$ directly. Its evaluation takes approximately 30 seconds per set of sketches on average.

4.2 Test Data

We use two real-life datasets, both consisting of visual descriptors extracted from images. The first one is formed by 1 million *DeCAF* [4, 15] descriptors from the *Profiset collection*[3]. These descriptors are 4,096-dimensional vectors of

[3] http://disa.fi.muni.cz/profiset/



(a) GHP to set values in one bit

(b) Two GHPs to set two bits of sketches

Fig. 4: Generalized hyperplane partitioning for sketching technique

float numbers taken as an output from the last hidden layer of a deep convolutional neural net [8]. Although this neural net has been trained for ImageNet classification, the last hidden layer is suitable for solving recognition problems [4, 15]. These descriptors with Euclidean distance L_2 form the metric space (D, d). The distance density is depicted in Figure 3a and the intrinsic dimensionality of dataset (defined in Section 2.2) is 26.9.

The second dataset is formed by 1 million *SIFT* descriptors [9] from ANN dataset[4]. These descriptors are 128-dimensional vectors of unsigned integers. We compare them with Euclidean distance as well, and density of this distance is depicted in Figure 3b. The intrinsic dimensionality of this dataset is 13.4.

4.3 Sketching Techniques

We examine four sketching techniques in this paper. The *GHP_50* technique is adopted from paper [12]. It is based on *generalized hyperplane partitioning* (GHP) depicted in Figure 4. A pair of pivots $p_{i1}, p_{i2} \in D$ is selected for each bit i of sketches $sk(o), o \in X$, and value of bit i expresses which of these two pivots is closer to o. Therefore, one instance of GHP determines one bit of all sketches $sk(o), o \in X$. The pivot pairs are selected to produce balanced and low correlated bits.

In particular, the pivot selection [12] works as follows: (1) an initial set of pivots P_{sup} is selected at random from domain D, (2) balance of GHP is evaluated using a sample set of X for all pivot pairs $(p_1, p_2), p_1, p_2 \in P_{sup}$, (3) set P_{bal} is formed by all pivot pairs that divide the sample set into two parts balanced with tolerance 5 % (at least 45 % to 55 %) and corresponding sketches sk_{bal} with balanced bits are created. (4) The absolute value of Pearson correlation coefficient is evaluated for all pairs of bits of sketches sk_{bal} to form correlation matrix M, and (5) a heuristic is applied to select rows and columns of M, which form its sub-matrix with low values and size $\lambda \times \lambda$. (6) Finally, the pivot pairs which produce the corresponding low correlated bits define sketches $sk(o), o \in X$. A pseudo-code of this heuristic is available online[5].

The second technique *GHP_80* is similar to GHP_50, but the pivots P_{bal} are selected to produce bits balanced to $\beta = 0.8$. This sketching technique have

[4] http://corpus-texmex.irisa.fr/

[5] https://www.fi.muni.cz/ xmic/sketches/AlgSelectLowCorBits.pdf

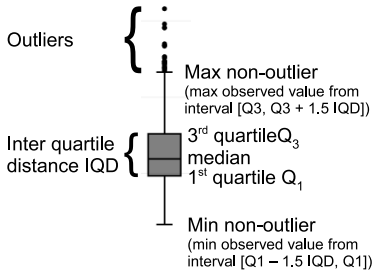


Fig. 5: Box-plot

been discussed to produce sketches of the similar quality as GHP_50, considering sufficiently high sketch length λ [13]. However, such sketches should be indexable more easily due to their lower intrinsic dimensionality.

Technique *BP_50* uses *ball partitioning* instead of GHP. BP is defined by a pivot and radius to split data into two parts. We evaluate distances to pivot for each object from a sample set to select radius dividing sample set into halves. Therefore λ pivots are selected to create sketches of length λ. To ensure as small pairwise bit correlations as possible, we employ the same heuristic as in case of techniques GHP_50 and GHP_80.

The last technique *THRR_50* is inspired by papers [10, 6], and it is the only one of examined sketching techniques which is applicable just in vector space, not in the more generic metric space. It uses the principal component analysis (PCA) to shorten vectors to length λ. Then these shortened vectors are rotated using a random matrix, and finally they are binarized using the median values of each their dimension: the bit value expresses, whether the value in a given position of rotated shortened vector is higher or lower then the median evaluated on a sample set. Therefore, the balance β of bits is 0.5. The random rotation of shortened vectors helps to distribute information equally over the vector, as the PCA returns vectors with decreasing importance of values in particular positions. Since the binarization dismiss this different importance, it is suitable to rotate shortened vectors randomly and then binarize [6].

We create sketches $sk(o), o \in X$ of four different lengths: 64, 128, 192 and 256 bits, by each of the sketching technique for the both, DeCAF and SIFT datasets. The only exception is constituted by THRR_50 on SIFT dataset, as this technique cannot produce sketches longer then the original vectors. Therefore, for this combination we examine just lengths 64 and 128 bits.

4.4 Results

To verify capability of Approaches A and PM to estimate quality of sketches, we compare these estimations with the recall of kNN queries, using the procedure and parameters described in Section 4.1. We use $CandSet(q)$ of size 2000 objects (i.e. 0.2 % of the dataset X), and we depict results by *box plots* to show distribution of values for particular query objects (see Figure 5 for its definition).

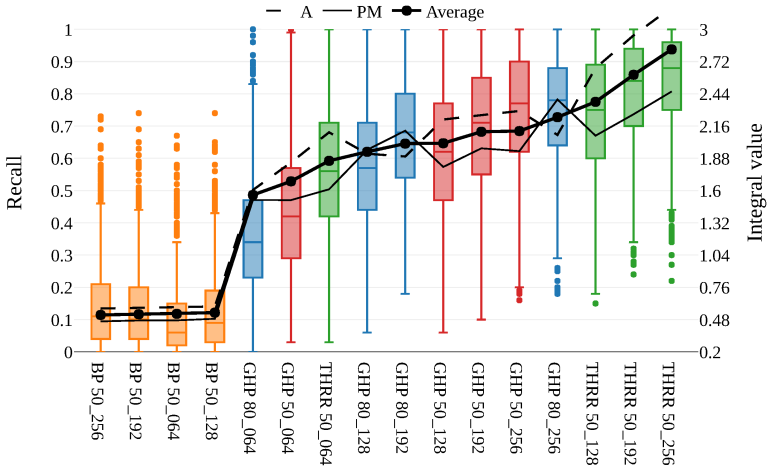


Fig. 6: DeCAF dataset: the measured recall and quality estimations by Approaches A, PM and their average

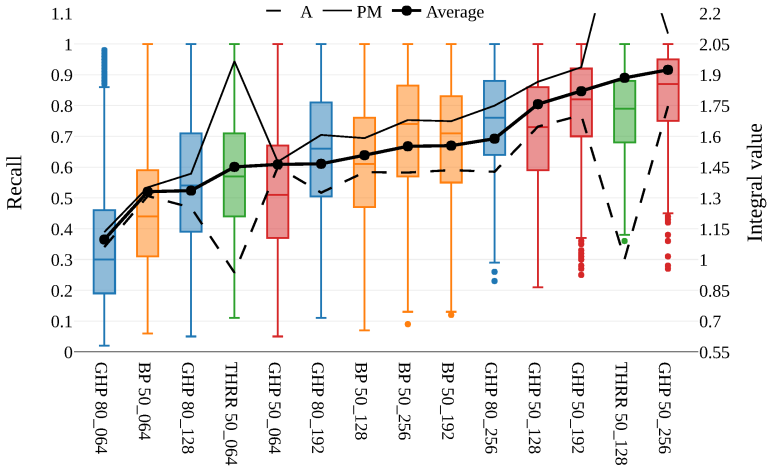


Fig. 7: SIFT: the recall, quality estimations by Approach A, PM and their average

The results for the DeCAF dataset are depicted in Figure 6. The box-plots, which describe the measured recall use the primary y axis. The names of particular sketching techniques (on x axis) are of the form $skTech\,\beta\_\lambda$ where β is the balance of bits in percentages and λ is the sketch length in bits. Quality estimations use the secondary axis y: Approach A is expressed by dashed line and Approach PM by full line. As we have two estimations of the same feature, we evaluate even the average of these two estimations, which is expressed by black curve with points.

The results for the SIFT dataset are depicted in the same way in Figure 7. There is a clear correspondence for both datasets between the recall and estima-

Table 2: Correlations of quality estimations and measured medians of the recall

	Approach A	Approach PM	Average
DeCAF	+0.96	+0.97	+0.98
SIFT	+0.55	+0.74	+0.93

tions. However, in case of SIFT, the Approach PM significantly overestimates sketches THRR and Approach A underestimates it. Another remark is, that the quality of sketching techniques is data dependent, as for instance the BP techniques perform bad on DeCAF dataset but for SIFT they are still reasonable. The interpretation of these results should take into account, that the recall is strongly dependent on the size of the $CandSet(q)$ as well.

We show the Pearson correlation coefficient between the estimated values and the medians of measured recalls in Table 2. The both approaches provide top quality results in case of DeCAF dataset, but the estimations are not as good in case of SIFT due to THRR technique. Nevertheless, the Approach PM still provides a strongly correlated estimation with the measured values, and the quality of averaged estimation is of a top quality even in case of this dataset.

We do not discuss an efficiency of query processing, as we consider sequential evaluation of all Hamming distances $h(sk(q), sk(o)), o \in X$ during query evaluation. In this case, query processing times are equal for all sets of sketches, as they are given by the size of candidate set. Indexing of sketches pays off mainly for huge datasets, and its efficiency is influenced by $iDim$ of sketches. However, our preliminary experiments on just two very different datasets do not justify any reasonable conclusions about this feature, and thus we postpone it to the future work.

5 Conclusions

Several techniques of the Hamming embedding have been proposed to speed up similarity search. Since their parameters (including the length of the transformed objects – sketches) must be selected in advance, and their ability to approximate similarity relationships between objects is data dependent, the selection of particular sketches for similarity search is a challenging problem.

We proposed two efficient approaches to estimate the quality of sketches with respect to particular data. These approaches do not need any ground truth or query evaluations but just a small sample set of data objects and their sketches. Both approaches are based on analytic study of sketches. Experiments with two real-life datasets show that they provide a reasonable estimation of the quality of sketches when compared with the recall of kNN queries. The average of the proposed estimations follow the medians of the measured recall with correlations +0.98 and +0.93, in cases of our two datasets.

Acknowledgements

Paper was supported by the Czech Science Foundation project GBP103/12/G084.

References

1. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada. pp. 380–388 (2002)
2. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Comput. Surv. 33(3), 273–321 (Sep 2001)
3. Daugman, J.: The importance of being random: statistical principles of iris recognition. Pattern recognition 36(2) (2003)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Icml. vol. 32, pp. 647–655 (2014)
5. Dong, W., Charikar, M., Li, K.: Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In: Proc. of the 31st annual int. ACM SIGIR conf. on Research and development in information retrieval. ACM (2008)
6. Gordo, A., Perronnin, F., Gong, Y., Lazebnik, S.: Asymmetric distances for binary embeddings. IEEE Trans. Pattern Anal. Mach. Intell. 36(1), 33–47 (2014)
7. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: 10th European Conference on Computer Vision (ECCV), Marseille, France,2008, Proceedings. pp. 304–317 (2008)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. vol. 60, pp. 84–90 (2017)
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV. pp. 1150–1157 (1999)
10. Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures. In: Proc. of the 2004 ACM CIKM Int. Conf. on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004. pp. 208–217 (2004)
11. Lv, Q., Josephson, W., Wang, Z., Charikar, M., Li, K.: Ferret: a toolkit for content-based similarity search of feature-rich data. ACM SIGOPS Operating Systems Review (2006)
12. Mic, V., Novak, D., Zezula, P.: Designing sketches for similarity filtering. In: IEEE International Conference on Data Mining Workshops, ICDMW 2016, December 12-15, 2016, Barcelona, Spain. pp. 655–662 (2016)
13. Mic, V., Novak, D., Zezula, P.: Sketches with unbalanced bits for similarity search. In: Similarity Search and Applications - 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings. pp. 53–63 (2017)
14. Muller-Molina, A.J., Shinohara, T.: Efficient similarity search by reducing i/o with compressed sketches. In: Proceedings of the 2nd Int. Workshop on Similarity Search and Applications. pp. 30–38 (2009)
15. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE CVPR conference (2014)
16. Samet, H.: Foundations of multidimensional and metric data structures. Morgan Kaufmann (2006)
17. Wang, Z., Dong, W., Josephson, W., Lv, Q., Charikar, M., Li, K.: Sizing sketches: a rank-based analysis for similarity search. In: Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2007, San Diego, California, USA, June 12-16, 2007. pp. 157–168 (2007), http://doi.acm.org/10.1145/1254882.1254900
18. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity search: the metric space approach, vol. 32. Springer (2006)